Erschienen in: Sprache und Datenverarbeitung Jg. 32 (2008) H. 1, S. 5-6.

Roman Schneider & Bernhard Schröder

Vorwort

Mit der rasanten Entwicklung einer ganzen Familie von XML-basierten texttechnologischen Standards seit mehr als zehn Jahren ist die Hoffnung verbunden, dass dadurch die technische Handhabung von linguistischen Daten enorm erleichtert werde und computerlinguistische Projekte davon in der Praxis stark profitierten. Dies um so mehr, als auch die meisten etablierten Datenbankmanagementsysteme (DBMS), Open Source-Lösungen sowie fachspezifische und/oder akademische Datenbankprodukte das Potential strukturorientierter Auszeichnungssprachen erkannt und Schritt für Schritt integriert haben. XML-Standards für die Modellierung der Mikrostruktur informationeller Angebote sowie für die Kommunikation mit externen Applikationen, kombiniert mit den bekannten Vorzügen einer datenbankgetützten Verwaltung und Weiterverarbeitung - aus dieser Perspektive heraus betrachtet sollte sich die computerlinguistische Forschung zukünftig vermehrt um das „Was?" und immer weniger um das oft leidige „Wie?" kümmern dürfen.

In der Tat gibt es für viele Probleme, die vor einigen Jahren noch umfangreiche Spezialentwicklungen erforderten, inzwischen sehr hilfreiche Standard-Werkzeuge. Doch gilt dies durchgängig? Und welche Überlegungen sind vor Implementierungsentscheidungen nach wie vor unerlässlich? Hier soll der interdisziplinäre GLDV-Workshop „Datenbanktechnologien für hypermediale linguistische Anwendungen", der im Rahmen der Konvens 2008 in Berlin stattfindet, für mehr Klarheit sorgen, indem er Erfahrungen mit hypermedialen Datenbanken aus unterschiedlichen texttechnologischen Projekten zusammenbringt. Die Vorträge und Systemdemonstrationen des Workshops liegen den Artikeln dieses Heftes zugrunde. Die Beiträge zeigen deutlich, dass nach wie vor nicht alle Werkzeuge alle naheliegenden Erwartungen erfüllen und jedes Projekt eine sorgfältige Bestandsaufnahme der verfügbaren technischen Lösungen erfordert. Die Berichte kommen aus höchst unterschiedlichen Anwendungsbereichen. Das Spektrum umfasst beispielsweise sowohl die Verwaltung unterschiedlicher Annotationsformen für geschrieben- wie für gesprochensprachliche Korpora; weiterhin werden Schnittstellen zu Online-Wörtebüchern und Wikis sowie die Verwaltung fachterminologischer Ontologien thematisiert.

Richard Eckart befasst sich mit den Kriterien, die bei der Auswahl einer nativen XML-Datenbank für die Verwaltung XML-annotierter Korpora anzulegen sind. Er zeigt, dass nicht alle Erwartungen, die man mit dem Einsatz von XML-Standards verbindet, gleichermaßen von den auf dem Markt befindlichen Datenbanksystemen eingelöst werden. Vor diesem Hintergrund diskutiert er die Vor- und Nachteile dreier einschlägiger Produkte.

Um Korpora gesprochener Sprache geht es bei Joachim Gasch. Er beschreibt die Herausforderungen, die sich aus der Integration unterschiedlicher Metadatensätze für gesprochensprachliche Korpora am Institut für deutsche Sprache (IDS) Mannheim ergeben. Unter Verwendung eines XML-fähigen objekt-relationalen Datenbankmanagementsystems illustriert er die Implementierung eines für Speicherung und Retrieval optimierten XML-Schema-Ansatzes.

Roman Schneider zeigt am Beispiel der Integration eines gedruckten Valenzwörterbuchs in ein sprachwissenschaftliches Webportal die aktuellen Möglichkeiten datenbankbasierter XML-Verarbeitung für die praktische Lexikographie auf. Neben verschiedenen Arten der Speicherung, Segmentierung, Auswertung und Transformation werden insbesondere das Retrieval mit Hilfe von SQL und XPath sowie die vermittels AJAX realisierte Anbindung an Web-Services thematisiert.

Alexander Mehler und andere befassen sich mit Wiki-Daten. Sie stellen ein API für die Auswertung verschiedener struktureller, Artikel-interner, sowie die Verlinkungsstruktur berücksichtigender (Meta-)Informationen vor. So lassen sich beispielsweise Link-Topologien zwischen Artikeln und Themengebieten und die Verknüpfungen, die sich durch Koautorenschaften bzw. Autoren-Cluster ergeben, ermitteln. Derartige Daten können Aufschluss über implizite Ontologien, die Relevanz von Beiträgen oder Autorengemeinschaften geben.

Ineta Sejane beschreibt die Entwicklung und datenbankbasierte Verwaltung einer domänenspezifischen Ontologie zur grammatischen Fachterminologie. Terminologische Basis und praktisches Anwendungsfeld ist das grammatische Web-Informationssystem GRAMMIS. Die Ontologie soll Navigation und Retrieval in Online-Angeboten vereinfachen, indem konkurrierende Termini unterschiedlicher theoretischer Grammatikansätze zueinander in Beziehung gesetzt werden.

Wir hoffen, dass das Spektrum der hier dargestellter Anwendungen einen gewissen Querschnitt durch die Praxis der Datenbanknutzung im texttechnologischen Kontext bieten, und wünschen eine anregende Lektüre.